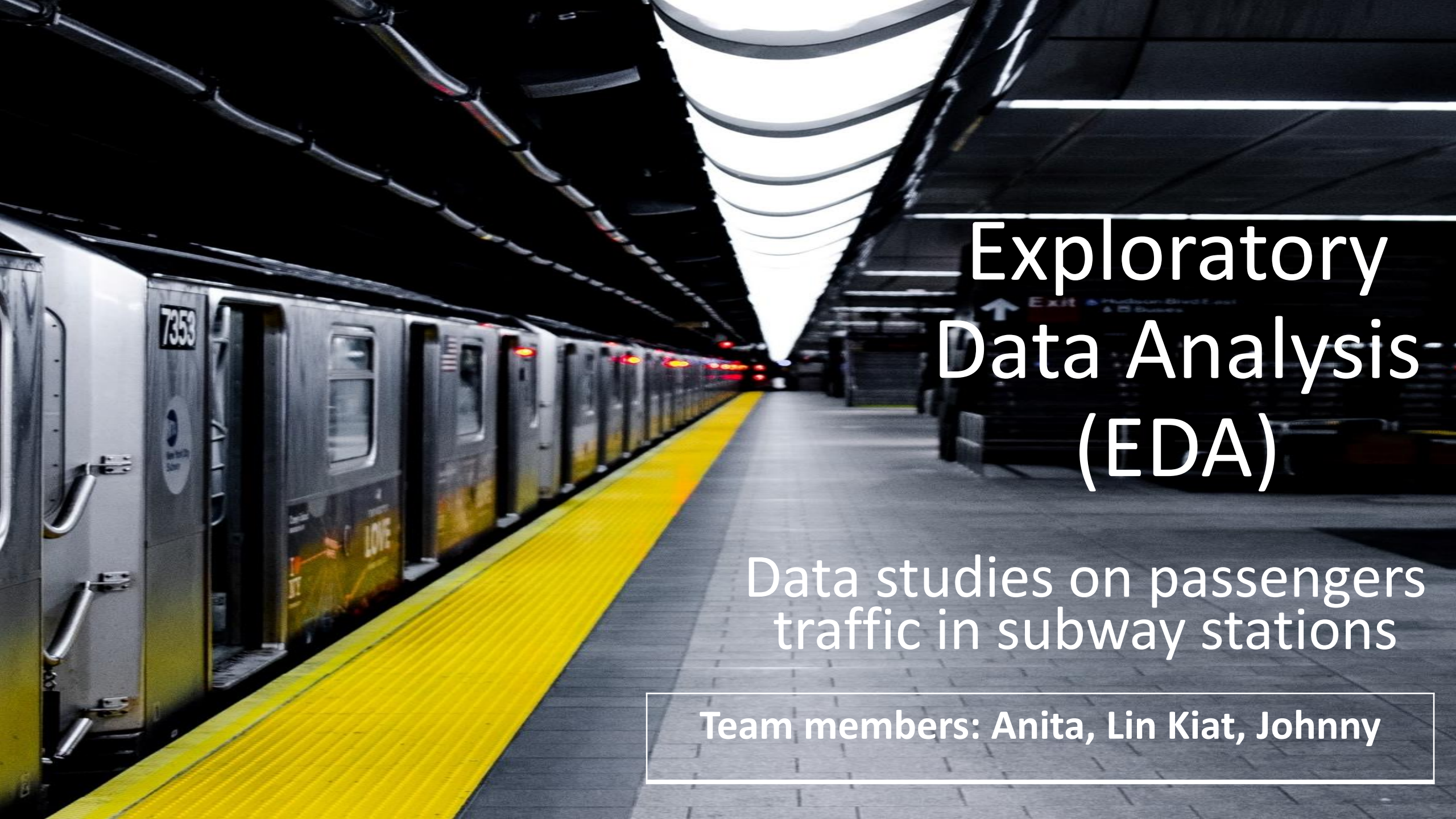


Exploratory Data Analysis (EDA)

Data studies on passengers traffic in subway stations

Team members: Anita, Lin Kiat, Johnny

Agenda

- Purpose.
- Quick facts on New York.
- Data management.
- Interpretation of data.
- Conclusion.

Purpose: To build awareness of Women Tech Women Yes (WTWY) International

Purpose: To build awareness of Women Tech Women Yes (WTWY) International

Strategy

- Put street teams at entrances of subway stations.

Purpose: To build awareness of Women Tech Women Yes (WTWY) International

Strategy

- Put street teams at entrances of subway stations.
- To collect email addresses from passengers and send email invites to them to attend gala dinner.

Purpose: To build awareness of Women Tech Women Yes (WTWY) International

Strategy

- Put street teams at entrances of subway stations.
- To collect email addresses from passengers and send email invites to them to attend gala dinner.
- To achieve maximum effect, the study will focus on putting teams at high traffic subway stations.

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m.

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m.

Persons between 19 to 64 yrs old: 56.6%.

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m.

Persons between 19 to 64 yrs old: 56.6%.

Females: 51.4%.

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m

Persons between 19 to 64 yrs old: 56.6%

Females: 51.4%

Households with internet subscription: 80.9%.

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m

Persons between 19 to 64 yrs old: 56.6%

Females: 51.4%

Households with internet subscription: 80.9%.

Mean travelling time on subway = 33.3min

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m

Persons between 19 to 64 yrs old: 56.6%

Females: 51.4%

Households with internet subscription: 80.9%.

Mean travelling time on subway = 33.3min

Target audience =

Quick facts on New York City.

(Source: US census bureau. 2019 data)

Population: 19.5m

Persons between 19 to 64 yrs old: 56.6%

Females: 51.4%

Households with internet subscription: 80.9%.

Mean travelling time on subway = 33.3min

Target audience = 19.5 x 56.6% x 51.4% x 80.9% = 4.6m

Data management

Data management

Issues encountered

Data management

Issues encountered

- Error in data source. E.g “EXITS” column was not format properly.

Data management

Issues encountered

- Error in data source. E.g “EXITS” column was not format properly.

```
for col in df.columns:
    print(col,len(col))
```

```
C/A 3
UNIT 4
SCP 3
STATION 7
LINENAME 8
DIVISION 8
DATE 4
TIME 4
DESC 4
ENTRIES 7
EXITS                                                                68
```

Check column length reveals blanks in "EXITS.

Data management

Issues encountered

- Error in data source. E.g “EXITS” column was not format properly.

```
for col in df.columns:
    print(col,len(col))
```

```
C/A 3
UNIT 4
SCP 3
STATION 7
LINENAME 8
DIVISION 8
DATE 4
TIME 4
DESC 4
ENTRIES 7
EXITS
```

Check column length reveals blanks in "EXITS.

Excess blank spaces in exits

Data management

Issues encountered

- Error in data source. E.g “EXITS” column was not format properly.

Solution:

Strip the excess blanks and make changes permanent.

```
for col in df.columns:
    df.rename(columns={col:col.strip()}, inplace=True)
```

Data management

Issues encountered

- Duplication of data.

Data management

Issues encountered

- Duplication of data.

```
(df
 .groupby(["C/A", "UNIT", "SCP", "STATION", "DATE_TIME"])
 .ENTRIES.count()
 .reset_index()
 .sort_values("ENTRIES", ascending=False)).head(5)
```

	C/A	UNIT	SCP	STATION	DATE_TIME	ENTRIES
87780	N333A	R141	00-00-00	FOREST HILLS 71	2020-08-26 05:00:00	2
40591	J023	R436	00-00-00	NORWOOD AV	2020-08-25 05:00:00	2
0	A002	R051	02-00-00	59 ST	2020-08-22 00:00:00	1
145213	R127	R105	00-00-02	14 ST	2020-08-28 09:50:21	1
145215	R127	R105	00-00-02	14 ST	2020-08-28 16:00:00	1

Shows there is duplicate entries in "FOREST HILLS" AND "NORWOOD AV"

Data management

Issues encountered

- Duplication of data.

```
(df
 .groupby(["C/A", "UNIT", "SCP", "STATION", "DATE_TIME"])
 .ENTRIES.count()
 .reset_index()
 .sort_values("ENTRIES", ascending=False)).head(5)
```

	C/A	UNIT	SCP	STATION	DATE_TIME	ENTRIES
87780	N333A	R141	00-00-00	FOREST HILLS 71	2020-08-26 05:00:00	2
40591	J023	R436	00-00-00	NORWOOD AV	2020-08-25 05:00:00	2
0	A002	R051	02-00-00	59 ST	2020-08-22 00:00:00	1
145213	R127	R105	00-00-02	14 ST	2020-08-28 09:50:21	1
145215	R127	R105	00-00-02	14 ST	2020-08-28 16:00:00	1

Shows there is duplicate entries in "FOREST HILLS" AND "NORWOOD AV"

Data management

Issues encountered

- Duplication of data.

```
(df
 .groupby(["C/A", "UNIT", "SCP", "STATION", "DATE_TIME"])
 .ENTRIES.count()
 .reset_index()
 .sort_values("ENTRIES", ascending=False)).head(5)
```

	C/A	UNIT	SCP	STATION	DATE_TIME	ENTRIES
87780	N333A	R141	00-00-00	FOREST HILLS 71	2020-08-26 05:00:00	2
40591	J023	R436	00-00-00	NORWOOD AV	2020-08-25 05:00:00	2
0	A002	R051	02-00-00	59 ST	2020-08-22 00:00:00	1
145213	R127	R105	00-00-02	14 ST	2020-08-28 09:50:21	1
145215	R127	R105	00-00-02	14 ST	2020-08-28 16:00:00	1

Shows there is duplicate entries in "FOREST HILLS" AND "NORWOOD AV"

Run data based on Date_Time format. Allows the count on duplication of values.

- Forest Hills: 2
- Norwood : 2

Data management

Issues encountered

- Duplication of data.

Solution

Drop duplicates, create a subset and make changes permanent

```
df.sort_values(["C/A", "UNIT", "SCP", "STATION", "DATE_TIME"],
                            inplace=True, ascending=False)
df.drop_duplicates(subset=["C/A", "UNIT", "SCP", "STATION", "DATE_TIME"], inplace=True)
```

Data management

Issues encountered

- 1 week of data. It is a sample data.

Data management

Issues encountered

- 1 week of data. It is a sample data.

a) May not be representative of the actual monthly data.

Data management

Issues encountered

- 1 week of data. It is a sample data.

a) May not be representative of the actual monthly data.

b) Traffic can be affected by factors such as train breakdown, events etc.

Data management

Issues encountered

- 1 week of data. It is a sample data.

a) May not be representative of the actual monthly data.

b) Traffic can be affected by factors such as train breakdown, events etc.

Solution

- Includes more weeks in the data.
- Exclude any outliners in the data.

Data management

Issues encountered

- Counter recording passenger traffic resets, once it exceeds a certain number.

Data management

Issues encountered

- Counter recording passenger traffic resets, once it exceeds a certain number.

```
df_daily[df_daily["ENTRIES"] < df_daily["PREV_ENTRIES"]].head()
```

	C/A	UNIT	SCP	STATION	DATE	ENTRIES	PREV_DATE	PREV_ENTRIES
267	A011	R080	01-03-00	57 ST-7 AV	08/23/2020	885655164	08/22/2020	885655237.0
268	A011	R080	01-03-00	57 ST-7 AV	08/24/2020	885655029	08/23/2020	885655164.0
269	A011	R080	01-03-00	57 ST-7 AV	08/25/2020	885654873	08/24/2020	885655029.0
270	A011	R080	01-03-00	57 ST-7 AV	08/26/2020	885654713	08/25/2020	885654873.0
271	A011	R080	01-03-00	57 ST-7 AV	08/27/2020	885654553	08/26/2020	885654713.0

Checking for current date entries is less than previous day entries.

Previous day entries is more than current date entries.

Data management

Issues encountered

- Counter recording passenger traffic resets, once it exceeds a certain number. Some data recorded in reverse order.

Solution

a) If current day entries < previous day entries.

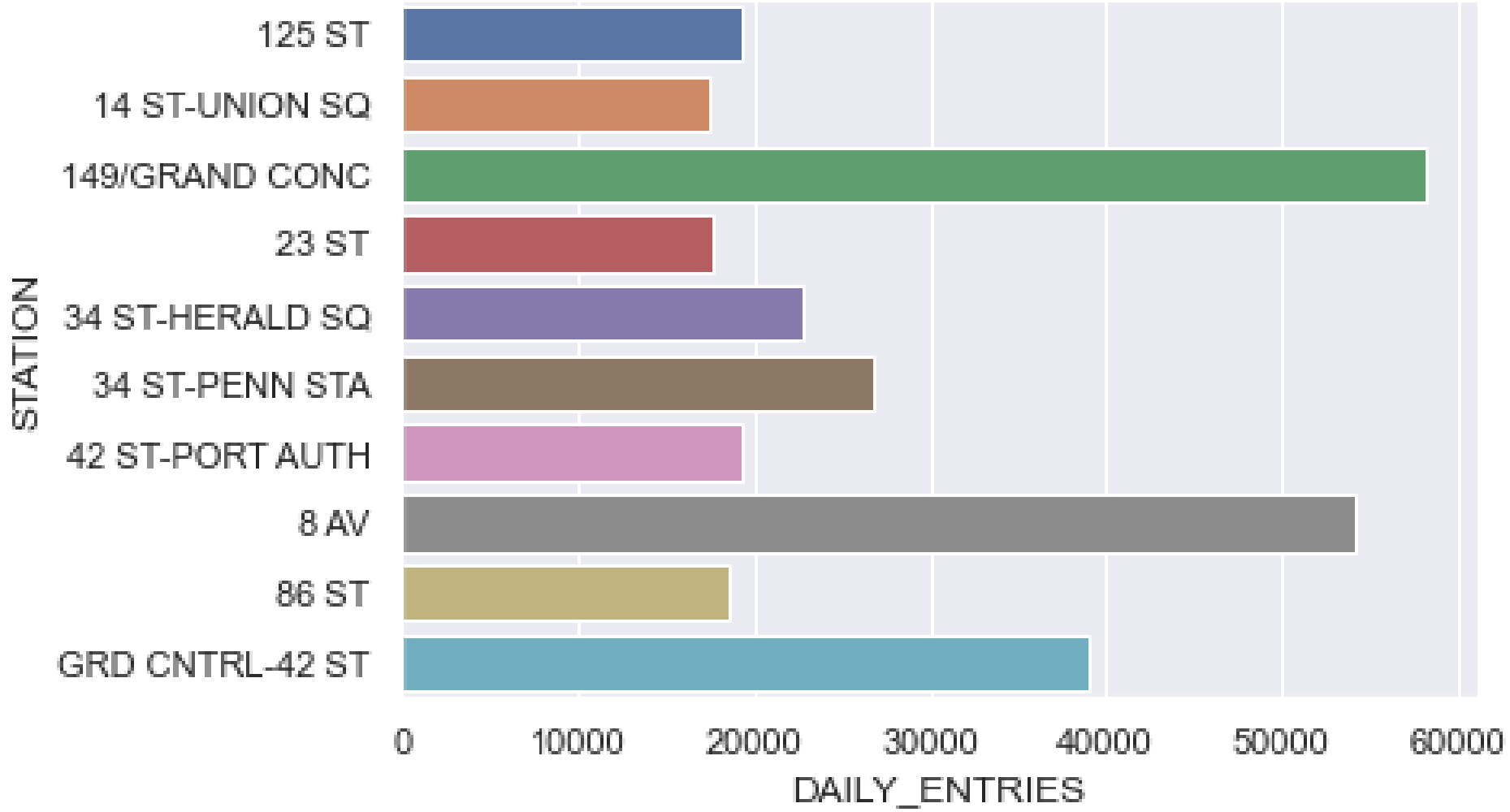
Ans. Minus (current day entries – previous day entries)

b) If current day entries > max counter.

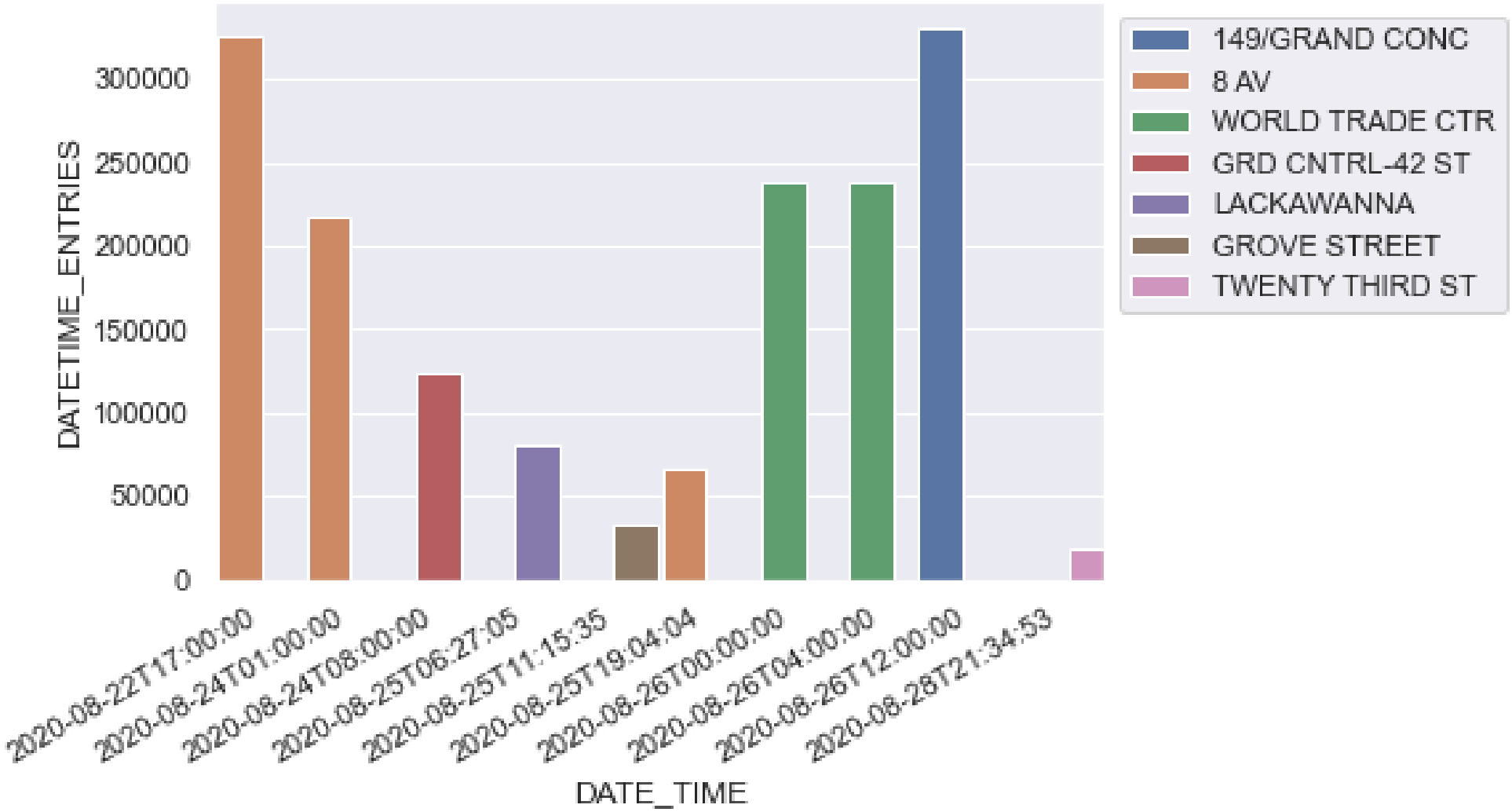
Ans. Take the min entries from either current day/previous day.

Interpretion of data

Top 10 stations in New York City



Top 10 stations by date time periods



Conclusion

- To allocate team to the top 3 passenger traffic stations and on high traffic time periods stations.
- To collect more weekly data to have a more accurate data on traffic.
- To investigate outliners to have better interpretation on data.
- To build a model based on EDA to predict better traffic flow.